动 物 学 研 究 2011，Dec. 32(6): 585−595 CN 53-1040/Q ISSN 0254-5853
Zoological Research DOI：10.3724/SP.J.1141.2011.06585

黑腹果蝇中嵌合新基因的进化命运和表达模式

占祖兵[1,2]，张 越[1]，赵若苹[1]，王 文[1,*]

(1. 中国科学院昆明动物研究所 遗传资源与进化国家重点实验室，云南 昆明 650223; 2. 中国科学院研究生院，北京 100049)

摘要：新基因的起源和进化对基因组多样性的产生具有重要的贡献。新基因起源常常通过外显子重排而形成嵌合的基因结构，以产生具有新功能的蛋白质。该文调查了在黑腹果蝇中的 14 个新起源的嵌合基因在群体中的多态性，发现其中 8 个在群体中的核苷酸多态性会引起提前终止子，而其他 6 个在群体中编码框都完整且其中 4 个受到负选择。研究结果表明，嵌合新基因起源后可能存在两种命运：积累提前终止子突变而假基因化，或者表现出一定功能而受自然选择固定下来。基因表达的数据显示，与 RNA 介导外显子重排(逆转座)形成的新基因不一样，这些由 DNA 水平外显子重排产生的新基因没有精巢或者雄性特异性表达模式，而是表现出更为多样性的时空表达模式，这提示尽管通过DNA水平外显子重排产生的新基因可能正在变成假基因或者非蛋白质编码的RNA基因，但它们依然可能具有进化出广泛的生物学功能的潜力。

关键词：黑腹果蝇；嵌合新基因；外显子重排；表达模式
中图分类号：Q969.462.2; Q344 **文献标志码**：A **文章编号**：0254-5853-(2011)06-0585-11

Evolutionary fate and expression patterns of chimeric new genes in *Drosophila melanogaster*

ZHAN Zu-Bing [1,2], ZHANG Yue [1], ZHAO Ruo-Ping [1], WANG Wen [1,*]

(1. *State Key Laboratory of Genetic Resources and Evolution*, *Kunming Institute of Zoology*, *the Chinese Academy of Sciences*, *Kunming* 650223, *China*; 2. *Graduate University of the Chinese Academy of Sciences*, *Beijing* 100049, *China*)

Abstract: Origin and evolution of new genes contribute a lot to genome diversity. New genes usually form chimeric gene structures through DNA-based exon shuffling and generate proteins with novel functions. We investigated polymorphism of 14 chimeric new genes in *Drosophila melanogaster* populations and found that eight have premature stop codons in some individuals while six are intact in the population, four of which are under negative selection, suggesting the two evolutionary fates of new chimeric genes after origination: accumulate premature stop codons and pseudolize, or acquire functions and get fixed by natural selection. Different from new genes originated through RNA-based duplication (retroposition) which are usually testis-specific or male-specific expressed, the expression patterns of these new genes through DNA-based exon shuffling are temporally and spatially diverse, implying that they may have the potential to evolve various biological functions despite that they may become pseudogenes or non-protein-coding RNA genes.

Key words *Drosophila melanogaster*; Chimeric new genes; Exon shuffling; Expression pattern

进化生物学上所说的新基因(new genes)，又称为年轻基因(young genes)，是指一个物种中基因组上新近起源的基因。从黑腹果蝇亚群(*Drosophila melanogaster* subgroup)中第一个新基因 *jingwei* (Long & Langley, 1993)的发现至今，新基因起源的分子机制在近二十年内得到了广泛的研究。这些分子机制包括基因重复(gene duplication)、外显子重排(exon shuffling)、逆转座(retroposition)、水平基因转移(lateral gene transfer)、基因分裂与融合(gene fusion/fission)、从头起源(*de novo* origination) (Long et al, 2003; Li et al, 2004; Zhou & Wang, 2008)。新基因起源后在进化上往往会有 3 种命运：亚功能化(subfunctionalization)、新功能化(neofunctionalization)和无功能化(nonfunctionallization) (Force et al, 1999;

收稿日期：2011-07-14；接受日期：2011-10-10
基金项目：国家自然科学基金重点项目(30930056)
[*]通讯作者(Corresponding author)，博士生导师, E-mail: wwang@mail.kiz.ac.cn

Lynch & Conery, 2000)。亚功能化和新功能化的新基因会在基因组上被保留下来，而无功能化的新基因往往会被假基因化而逐渐消失。新基因假基因化有两种途径：(1)积累有害突变而破坏读码框(ORFs, open reading frames)，如插入/缺失突变(indel, insertion/ deletion)、移码突变等; (2)失去转录功能而不表达。以前的案例研究发现，新基因起源常常形成嵌合的基因结构，如果蝇中的 *jingwei*、*sphinx*、*Adh–Twain* (Long & Langley, 1993; Wang et al, 2002; Jones & Begun, 2005)，以及人中的 *PIP5K1A*、*PMCHL1*、*PMCHL2* (Courseaux & Nahon, 2001; Babushok et al, 2007)。在黑腹果蝇中，约 30%新基因形成嵌合的结构，而嵌合的序列可来源于其它基因、转座子、简单重复序列或者基因间的非编码区(Zhou et al, 2008)。

嵌合新基因可以通过 RNA 介导外显子重排(逆转座)和 DNA 水平的外显子重排形成。前者通过将一个基因的 mRNA 反转录成 cDNA 后插入到基因组中，并招募其他基因的外显子，从而形成具有嵌和结构的新基因，如此前报道的 *jingwei* (Long & Langley, 1993; Long et al, 1999; Wang et al, 2000)(图 1A)。此前关于逆转座形成的嵌合基因在果蝇、植物、小鼠和人等不同物种中得到了较为广泛而系统的研究(Betrán et al, 2002; Emerson et al, 2004; Wang et al, 2006; Bai et al, 2007)。其中有不少关于逆转座形成的嵌合基因的经典案例，如果蝇中的 *jingwei*、*sphinx*、*Adh–Twain* (Long & Langley, 1993; Wang et al, 2002; Jones & Begun, 2005)，哺乳动物中的 *PIP5K1A–PSMD4*、*TRIM5–CypA*、*RBMXL1*、*Utp14c* (Sayah et al, 2004; Marques et al, 2005; Rohozinski et al, 2006; Babushok et al, 2007; Brennan et al, 2008; Virgen et al, 2008; Wilson et al, 2008)。DNA 水平的外显子重排则是通过 DNA 水平的重复事件，如基因重复(gene duplication)、部分基因重复(partial gene duplication)、片段重复(segment duplication)等，将两个或者多个基因的外显子融合到一起，或者基因内部的外显子产生重复而形成嵌合的基因结构(Long et al, 2003; Li et al, 2004; Zhou & Wang, 2008)，如图 1B。通常通过内含子介导的重组或者异常重组(illegitimate recombination)，将这两个或者多个部分融合到一起形成新基因(Gilbert, 1987; van Rijk et al, 1999)。最近在黑腹果蝇亚群中的研究显示，黑腹果蝇(*Drosophila melanogaster*) 中共有 14 个在 DNA 水平外显子重排产生的嵌合基因(Rogers et al, 2009)，其中 8 个是 *D. melanogaster* 特有的，即起源于 540 万年前(表 1)。嵌合基因以约 11.4 个基因/百万年的速度产生，随后以相接近的速度消亡，其中仅有 1.4%在基因组上被固定下来。

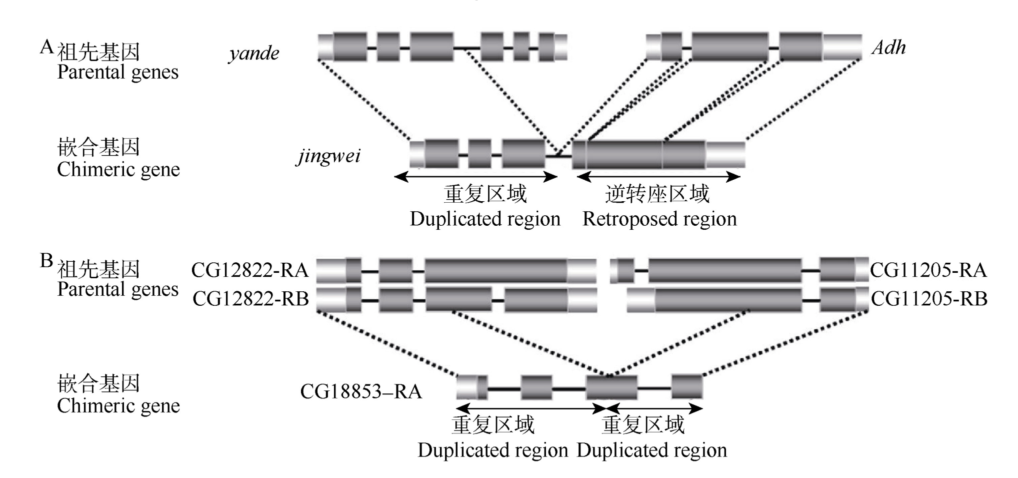


图 1 通过外显子重排产生嵌合基因的两种机制

Fig. 1 Two molecular mechanisms of chimeric gene formation by exon shuffling

A：RNA 介导的外显子重排(逆转座)：一个基因(*Adh*)通过逆转录后插入到另外一个基因(*yande*)内部，两个基因的外显子一起转录产生嵌合的新基因(*jingwei*); B：DNA 水平外显子重排：基因组上通过重复事件而将两个基因(*CG12822* 和 *CG11205*)的一部分拉到一起形成嵌合的新基因(*CG18853*)。基因的重复区域和逆转座区域如图中所示。

A: RNA-based exon shuffling (retroposition), in which the reverse transcribed copy of one gene (*Adh*) was inserted into another gene (*yande*), and the exons of these two genes were transcribed together and formed a young chimeric gene (*jingwei*); B: DNA-based exon shuffling, in which the parts of two parental genes (*CG12822*and *CG11205*) were duplicated and transcribed together and formed a young chimeric gene (*CG18853*). Duplicated regions and retroposed regions were illustrated in the figure.

表 1 嵌合新基因起源的年代和所在的物种

Tab. 1 The ages and species of the chimeric new genes

新基因 New gene	祖先基因 A [a] Par gene A [a]	祖先基因 B [a] Par gene B [a]	年龄 [b] Age (Myr) [b]	物种 [c] Species [c]
CG31904	*CG13796*	*CG7216*	＜5.4	Dmel
CG18853	*CG12822*	*CG11205*		
CG32318	*CG9191*	*CG9187*		
CG31864	*CG12264*	*CG5202*		
CG12592	*CG18545*	*CG12819*		
CG31687	*CG2508*	*CG31688*		
CR18217	*CG17286*	*CG4098*		
CG31668	*CG33124*	*CG8451*		
CG6653	*CG31002*	*CG17200*	5.4～12.8	Dmel, Dsim, Dsec, Dyak, Dere
CG17196	*CG17197*	*CG17195*		
CG30457	*CG10953*	*CG13705*	12.8～44.2	Dmel, Dsim, Dsec, Dyak, Dere, Dana
CG11961	*CG9416*	*CG30049*	44.2～54.9	Dmel, Dsim, Dsec, Dyak, Dere, Dana, Dspe
CG3978	*CG9656*	*CG10278*		
CG6844	*CG5610*	*CG11348*		

[a] Par gene A：祖先基因 A；Par gene B：祖先基因 B；[b] 年龄：嵌合基因起源至今的年龄； Myr：百万年； [c] 物种名称简写(Dmel：*Drosophila melanogaster*； Dsim： *D. simulans*；Dsec： *D. sechellia*；Dyak： *D. yakuba*；Dere： *D. erect*； Dana： *D. ananassae*；Dspe：*D. pseudoobscura*)。这些数据来源于 Rogers et al (2009)的报道。

[a] Par gene A denotes parental gene A, while Par gene B denotes parental gene B; [b] Age denote the ages of these chimeric new genes. Myr denotes million years. [c] The abbreviation of species names (Dmel: *Drosophila melanogaster*, Dsim: *D. simulans*, Dsec: *D. sechellia*, Dyak: *D. yakuba*, Dere: *D. erecta*, Dana: *D. ananassae*, Dspe: *D. pseudoobscura*). This data was derived from previous report by Rogers et al (2009).

由逆转座形成的嵌合基因常常表现为雄性或者精巢特异性的表达模式，这可能在雄性生殖系统的进化中扮演着重要的角色(Betrán et al, 2002; Emerson et al, 2004)。由外显子重排形成的嵌合基因，具有怎样的表达格局，还不清楚。

我们分析了这 14 个由外显子重排而产生的嵌合新基因(Rogers et al, 2009)在 *D. melanogaster* 群体中多态性的分布，根据其中多态性位点上核苷酸的替换对其编码区的影响而初步判断其是否有功能，如编码区的核苷酸替换是否会引起提前终止子(premature stop codons)，它们是否显著地正选择(positive selection)或者负选择(negative selection)。这些分析有助于更好地了解这些嵌合新基因起源后在进化上的命运，即是它们是以假基因(pseudogene)的形式退出进化舞台，还是继续进化出新的功能。通过分析其表达模式，我们观察到这些嵌合基因中大部分与已经报道的案例不同，它们并不像逆转座新基因那样多呈现雄性或者精巢特异性的表达，而是呈现更为多样性的表达模式。

1 材料与方法

1.1 研究材料

本文研究的 14 个嵌合新基因来源于此前的报道(Rogers et al, 2009)，这些基因起源的年代和所在的物种以及其祖先基因见表 1。这些新基因起源于不同的年代，其中 8 个是 *D. melanogaster* 特有的(年龄小于 540 万年)。

我们从 Ensembl Genome Browser (http://www.ensembl.org/)下载到这 14 个新基因以及其祖先基因的蛋白质编码序列、蛋白质的氨基酸序列以及其对应的序列号码，其对应的基因组版本为 Ensembl Genes 61、BDGP5.25。同时，我们还从该网站下载到了 *D. melanogaster* 的全部多态性数据，这些多态性数据来源于果蝇群体基因组学项目网站 DPGP (Drosophila Population Genomics Project: http://www.dpgp.org/)，Ensembl 网站已经将相应的序列比对到 *D. melanogaster* 测序参考基因组中基因区域(包括蛋白质编码区域(protein-coding regions)、非翻译的转录区域(UTRs, untranslational regions)、内含子区域(intronic regions)、以及基因间区域)，其对应的版本为 Ensembl Variation 61。这些多态性数据分别来源于 50 个 *D. melanogaster* 品系，即 MW11-1_1、MW27-3_1、MW28-1_1、MW28-2-3_1、MW38-1_1、MW38-2_1、MW46-1_1、MW56-2-3_1、MW6-1_1、MW6-2_1、MW6-3_1、MW63-1_1、MW63-2-3_2、MW9-1_1、MW9-2_1、RAL-301_1、

RAL-303_1、RAL-304_1、RAL-306_1、RAL-307_2、RAL-313_1、RAL-315_1、RAL-324_1、RAL-335_2、RAL-357_1、RAL-358_1、RAL-360_1、RAL-362_2、RAL-365_1、RAL-375_1、RAL-379_1、RAL-380_2、RAL-391_2、RAL-399_1、RAL-427_1、RAL-437_1、RAL-486_1、RAL-514_1、RAL-517_1、RAL-555_1、RAL-639_1、RAL-705_1、RAL-707、RAL-714_1、RAL-730_1、RAL-732_1、RAL-765_1、RAL-774_1、RAL-786_1、RAL-799_1、RAL-820_1 和 RAL-852_1，其中 MW 指果蝇来源于非洲东南部马拉维共和国(Malawi)，RAL 指果蝇来源于美国北卡罗来纳州首府罗莉(Raleigh,NC)。

1.2 中性检验与群体遗传学分析

为检测这些嵌合基因是否具有功能，我们使用多种遗传检验来判断这些新基因是否受到显著的正选择或者负选择。我们将从 Ensembl Variation 61 上下载的多态性数据还原到不同品系中的序列上，并抽提出其蛋白质编码序列用以进一步分析。这些多态性数据主要分为 3 类：同义核苷酸替换位点(synonymous substitution sites: SS)、错义核苷酸替换位点(non-synonymous substitution sites, NSS)和提前终止子突变(synonymous substitution sites, STOP)。首先，我们将含有提前终止子突变的基因分离出来，这些基因很可能正在假基因化(pseudogenization)，因此可能没有功能；其次，针对 *D. melanogaster* 特有的嵌合基因，由于这类基因太年轻且缺乏直系同源基因，其参考序列与祖先基因之间的替换数目太少，既无法在旁系同源基因进行检验，也无法在直系同源基因之间进行检验。因此，我们只能通过比较群体中多态性位点来判断这些嵌合基因是否受到选择。我们用 DnaSP 软件包(Rozas et al, 2003)分析这些嵌合基因的多态性，并计算 Tajima's D (Tajima, 1989)和 Fu-Li's D*/F* (Fu & Li, 1993; Fu, 1997)。我们还使用 MEGA 4.0 (Tamura et al, 2007)估算了这些嵌合基因在群体中平均每个同义/错义核苷酸位点上的替换数目(d_S和 d_N)以及 d_N/d_S。此外，对于其他物种有直系同源基因的嵌合基因，我们还使用 MEGA 4.0 (Tamura et al, 2007)估算了这些嵌合基因与其直系同源基因之间同义/错义替换率以及其比值(*Ka/Ks*)。

此外，针对假基因 *CR18217*，我们还进一步调查了 *D. melanogaster* 在全世界其他地区的不同品系中的多态性情况。调查的品系包括：CS、HG、OR、EC154、EC157、EC167、EC174、301A、303A、313A、315A、335A、350A、375A、732A、736A、740A、787A、799A 等。我们设计了两对 PCR 反应引物，以便扩增基因全长，引物序列分别为：(1)上游引物 F1：5′-*CGTTCGCACTGCAAACTGTAACT* -′3，下游引物 R1：5′-*TCACGTTACTTTCTGATTCGGGC*-′3；(2)上游引物 F2：5′-*GTTCGGAAAATATATGGAACATTG*-′3，下游引物 R2：5′-*ACCATTAGGCAGTTGATCTTAACTC*-′3。测序引物为以上 PCR 引物加上 5′-*GTAACAGACGACACCAT CGATC*-′3 (上游)，5′-*GACCGCTGTATGGCAACCATC*-′3(下游)。PCR 反应条件为：94 ℃ 3 min, 95 ℃ 30 s, 56 ℃ 30 s, 72 ℃ 1 min 30 s, 35 次循环, 72 ℃ 10 min, 4 ℃ 保存。 PCR 产物在 1.2%琼脂糖凝胶电泳下观察，并使用天根生化科技有限公司(TIANGEN)的琼脂糖凝胶回收试剂盒回收 PCR 产物，并用 BigDye 试剂盒进行测序，测序反应总体积 5 μL，包括 0.5 μL BigDye、0.5 μL 引物稀释液、1 μL 测序缓冲液、3 μL DNA 溶液和蒸馏水。测序反应条件为 96℃ 1min, 95 ℃ 10 s, 50 ℃ 5 s, 60℃ 4 min, 25 次循环, 4 ℃ 保存。G50 柱纯化测序反应产物，后甲酰胺溶解后置 3700 测序仪进行测序。产生的序列用 Lasergene 软件中的 SeqMan.exe 程序进行组装和分析。

1.3 基因表达分析

1.3.1 EST 分析 我们将从 NCBI 上下载的 *D. melanogaster* ESTs 序列(http://www.ncbi.nlm.nih.gov/nucest)用 BLAST 软件(Altschul et al, 1997)中的 Blastn 比对到这些嵌合基因以及其祖先基因的序列上。由于这些新产生的嵌合基因与其祖先基因在核苷酸序列上的同源性太高，我们采取了较为严格的标准，即选取那些相似度在 95%以上，比对长度占 EST 长度的 90%以上的序列，进一步去除那些同时比对到基因组上多个基因的 ESTs，仅保留能特异性的匹配到嵌合基因或者其祖先基因上的 ESTs。从 NCBI 上下载这些特异性的 ESTs 的组织或者发育时期的信息。由此，我们初步判断这些嵌合基因以及其祖先基因的表达组织。

1.3.2 高通量表达模式数据分析 从 FlyBase 下载由转录组测序产生的 *D. melanogaster* 高通量表达模式数据(ftp://flybase.org/flybase/associated_files/Gelbart.2010.10.13.tar.gz)。我们从中抽提出这 14 个嵌合基因以及其祖先基因表达的数据。这些表达数据分别来源于 30 个不同的发育时期：胚胎 00～02

h、02～04 h、04～06 h、06～08 h、08～10 h、10～12 h、12～14 h、14～16 h、16～18 h、18～20 h、20～22 h、22～24 h；幼虫 L1、L2、L3 各 12 h，L3 膨胀期(puffstage) 1～2 h、L3 膨胀期 3～6 h、L3 膨胀期 7～9 h；白色预蛹(white prepupae)前期、白色预蛹 12 h、白色预蛹 24 h，蛹 2 d、3 d、4 d；成虫雄性 1 d、5 d、30 d，成虫雌性 1 d、5 d、30 d。我们将这些发育时期分为 5 种, 即胚胎、幼虫、蛹、成虫雄性和雌性。根据 FlyBase 提供的标准, 表达量高低分为 9 个级别, 即不表达(0)、极度低表达(1～10)、表达量很低(11～100)、低表达(101～400)、中等表达(401～1400)、中等高度表达(1 401～4 000)、高表达(4 001～1 万)、表达量很高(1.0001 万～10 万、 极度高表达(10.0001 万～2 00 万)。

2 结 果

2.1 嵌合新基因的中性检验以及群体遗传学分析

本文所研究的 14 个嵌合新基因来源于 Rogers et al (2009)的报道(表 1)。这些嵌合新基因中有 6 个(*CG31904*、*CG31687*、*CG17196*、*CG11961*、*CG3978* 和 *CG6844*)存在多种选择性剪切形式, 其中 2 个(*CG31904* 和 *CG31687*)是新近起源的, 而另外 4 个则属于起源较早的嵌合基因; 28 个祖先基因中有 8 个存在多种选择性剪切形式。对于那些新近起源的年轻的基因, 由于起源年代很近, 在短期内单个品系无法积累很多突变, 根据仅有的基因组参考序列, 我们无法判断其与祖先基因之间的分歧以及其是否有功能而受到自然选择。因此, 我们利用 14 个新基因以及其祖先基因的在果蝇 50 个不同的品系或者个体的全部多态性数据, 提取出位于这些嵌合基因和其祖先基因中的蛋白质编码区域(protein-coding regions)的核苷酸替换数目以及类型。结果显示, 14 个嵌合基因中有 8 个(57.1%)在有些个体中积累了提前终止子突变(表 2)。没有积累提前终止子突变的6个基因中有4个是*D. melanogaster* 特有的, 即*CG32318, CG31864, CG12592* 和 *CG31687*。

进一步比较这 6 个没有积累提前终止子突变的嵌合基因在群体中的错义与同义替换率以及比值, Z 检验结果显示只有相对较古老的两个嵌合基因, 即 *CG30457* 和 *CG17196*, 受到负选择($d_N/d_S < 1$, $P < 0.05$)(表 3)。其余 4 个 Z 检验没有检测到选择信号的 *D. melanogaster* 特有的嵌合基因(*CG32318, CG31864*、*CG12592* 和 *CG31687*)中, *CG31687* 接近显著(d_N/d_S=0.507, P=0.08065)。Tajima's D 和 Fu-Li's D*/F*检验显示，*CG32318*、*CG12592* 和 *CG31687* 可能受到显著的自然选择(表 3)。其中, *CG12592* 和 *CG31687* 的 d_N/d_S<1 且 Tajima's D<0, Fu-Li's D*/F*<0, 表明它们受到显著的负选择。*CG32318* 的 d_N/d_S>1, 且 Tajima's D<0, Fu-Li's D*/F*<0, 表明该基因可能在群体内正在受到正选择作用。因此, 群体遗传学分析显示, 在 14 个嵌合基因中有 6 个(42.9%)不仅编码框在群体中完整而且不同程度受到负或正选择, 可能有功能。其余的 8 个基因在群体中积累了提前终止子突变。因此，它们可能是假基因或者是非蛋白编码的 RNA 基因。

2.2 嵌合新基因的表达分析

2.2.1　基于 EST 数据的表达分析　我们使用 EST 数据来调查这些嵌合新基因在哪些组织中表达。由于这些嵌合基因起源年代较近, 因此可能有部分 EST 能同时比对到两个或者两个以上的同源基因中。为准确区分新基因与其祖先基因的表达, 我们仅分析了那些能比对到唯一的基因上的 ESTs, 并以此分析嵌合基因与其祖先基因之间的表达组织。如表 4 所示, 14 个嵌合基因中有 7 个没有基因特异性的 EST 数据, 另外 7 个在头、胚胎、精巢、幼虫-蛹等组织或者时期表达。其中有 5 个基因在胚胎时期有表达, 2 个基因在头部表达, 仅有一个基因(*CG31684*)在成虫精巢中表达, 但由于 EST 数据有限, 无法确认是否是精巢特异性表达。假基因 *CR18217* 在胚胎时期表达, 其它 7 个积累了提前终止子突变的嵌合基因(可能正在假基因化)中有 3 个没有基因特异性的 EST 数据, 3 个在胚胎时期表达, 1 个在头部表达。28 个祖先基因中有 21 个有 EST 数据, 它们在多种组织中表达, 其中 6 个在精巢中表达。我们特别注意到可能发生假基因化的基因也表现出多样化的表达。EST 数据显示, *CG31904* 在头部表达, *CG31864* 在成虫精巢中表达, *CR18217*、*CG6653* 和 *CG3978* 在胚胎时期表达。尽管能特异性比对到嵌合基因的 EST 数据有限, 只有 7 个基因有基因特异性的 EST 数据, 但这些嵌合基因可能并不像逆转座新基因一样局限于精巢特异性表达, 而具有更为多样化的表达模式。

2.2.2　基于高通量表达模式数据的表达分析　为进一步分析这些嵌合基因的表达模式以推测可能在哪些发育时期执行功能, 我们分析了这些嵌合基因在果蝇生命周期中的 30 个时期的表达情况。在

表 2 嵌合基因与祖先基因在 *Drosophila melanogaster* 群体中的同义、错义替换和提前终止子的分布

Tab. 2 Distribution of synonymous, non-synonymous substitutions and premature stop codons of chimeric genes and their parental genes in *Drosophila melanogaster* populations

	基因名称(Gene ID)[a]			单碱基替换 (SNS)[b]		
	CG ID	FBgn ID	FBtr ID	SS	NSS	STOP
新基因 New gene	*CG31904*	*FBgn0260479*	FBtr0079501	29	45	1
			FBtr0079502	29	45	1
祖先基因 A Parental gene A	*CG13796*†	*FBgn0031939*	FBtr0079503	49	33	0
			FBtr0079504	49	33	0
			FBtr0079505	49	33	0
			FBtr0114496	49	33	0
祖先基因 B Parental gene B	*CG7216*	*FBgn0014454*	FBtr0079500	20	25	2
新基因	*CG18853*	*FBgn0042173*	FBtr0089426	2	9	1
祖先基因 A	*CG12822*†	*FBgn0033229*	FBtr0088841	7	11	0
			FBtr0088842	7	11	0
祖先基因 B	*CG11205*†	*FBgn0003082*	FBtr0088838	19	30	0
			FBtr0088839	18	29	0
新基因	***CG32318‡***	***FBgn0052318***	FBtr0072732	5	8	0
祖先基因 A	*CG9191*	*FBgn0004378*	FBtr0072733	60	54	2
祖先基因 B	*CG9187*†	*FBgn0035194*	FBtr0072731	15	8	0
新基因	***CG31864‡***	***FBgn0051864***	FBtr0080288	1	1	0
祖先基因 A	*CG12264*	*FBgn0032393*	FBtr0080290	15	25	1
祖先基因 B	*CG5202*†	*FBgn0032391*	FBtr0080286	8	15	0
			FBtr0100626	3	5	0
新基因	***CG12592‡***	***FBgn0037811***	FBtr0082233	8	17	0
祖先基因 A	*CG18545*†	*FBgn0037812*	FBtr0082234	2	7	0
祖先基因 B	*CG12819*	*FBgn0037810*	FBtr0082231	26	56	4
			FBtr0082232	26	56	4
新基因	***CG31687‡***	***FBgn0051687***	FBtr0081361	18	38	0
			FBtr0302224	9	15	0
祖先基因 A	*CG2508*	*FBgn0032863*	FBtr0081362	24	18	1
祖先基因 B	*CG31688*	*FBgn0051688*	FBtr0273378	13	12	1
			FBtr0273379	18	36	1
新基因	*CR18217**	*FBgn0036646*	FBtr0301925	NE	NE	NE
祖先基因 A	*CG17286*	*FBgn0027500*	FBtr0075363	20	49	1
祖先基因 B	*CG4098*†	*FBgn0036648*	FBtr0075361	0	12	0
新基因	*CG31668*	*FBgn0051668*	FBtr0113412	34	42	4
祖先基因 A	*CG33124*†	*FBgn0053124*	FBtr0300181	54	62	0
祖先基因 B	*CG8451*	*FBgn0031998*	FBtr0079637	33	16	1
新基因	*CG6653*	*FBgn0040255*	FBtr0082376	19	38	1
祖先基因 A	*CG31002*†	*FBgn0051002*	FBtr0085813	17	30	0
祖先基因 B	*CG17200*	*FBgn0040253*	FBtr0082377	16	31	2
新基因	***CG30457‡***	***FBgn0050457***	FBtr0086998	20	19	0
祖先基因 A	*CG10953*	*FBgn0034204*	FBtr0086997	22	13	1
祖先基因 B	*CG13705*	*FBgn0035582*	FBtr0073352	30	47	1
新基因	***CG17196‡***	***FBgn0039368***	FBtr0084922	19	6	0
			FBtr0302177	17	6	0
祖先基因 A	*CG17197*	*FBgn0039367*	FBtr0290204	19	29	1
祖先基因 B	*CG17195*†	*FBgn0039369*	FBtr0084921	12	10	0
新基因	*CG11961*	*FBgn0034436*	FBtr0086519	95	23	2
			FBtr0086520	96	28	2
祖先基因 A	*CG9416*†	*FBgn0034438*	FBtr0086555	86	19	0
祖先基因 B	*CG30049*	*FBgn0050049*	FBtr0087906	55	56	2
新基因	*CG3978*	*FBgn0003117*	FBtr0083220	40	48	1
			FBtr0083221	46	58	1
祖先基因 A	*CG9656*	*FBgn0001138*	FBtr0081808	20	48	2
			FBtr0300040	39	94	5
祖先基因 B	*CG10278*	*FBgn0038391*	FBtr0083218	44	72	2
新基因	*CG6844*	*FBgn0000039*	FBtr0084639	42	10	1
			FBtr0084640	42	10	1
祖先基因 A	*CG5610*	*FBgn0000036*	FBtr0084619	44	35	1
祖先基因 B	*CG11348*	*FBgn0000038*	FBtr0073299	45	12	0
			FBtr0073300	1	5	1

[a] 基因名称：CG ID 和 FBgn ID 分别是一个基因的两种不同名称；FBtr ID 是指一个基因的转录本名称。[b] 单碱基替换：SS 指编码区内核苷酸同义替换数目；NSS 指非同义替换(错义替换)数目；STOP 指核苷酸替换产生提前终止子数目； **CR18217*：*CR18217* 被注释成假基因(NE 指不存在); †指群体中没有提前终止子突变的祖先基因(11/28)； ‡指群体中没有提前终止子突变的新基因(6/14)。

[a] Gene ID: CG ID and FBgn ID denote two names of each new gene or parental gene; [b] SNS: SNP denotes single nucleotide substitution in protein-coding regions. SS denotes the number of synonymous substitutions, NSS denotes the number of non-synonymous substitutions, and STOP denotes the number of pre-mature stop codons; **CR18217*: *CR18217* was annotated as a pseudogene (NE: not existed); † The parental genes which do not have pre-mature stop codons in population (11/28); ‡ The new genes which do not have pre-mature stop codons in population (6/14).

表 3　嵌合新基因的中性检验结果

Tab. 3　Results of neutral tests on chimeric genes

CG 名称 [a] CG ID [a]	FBgn 名称 [a] FBgn ID [a]	FBtr 名称 [a] FBtr ID [a]	序列数 Seq *N*	位点数 Sites	d_N [b]	d_S [b]	d_N/d_S	*P* value [c]	Tajima's D	Fu-Li's D*	Fu-Li's F*
CG32318	***FBgn0052318***	FBtr0072732	21	2025	0.00554	0.00371	1.493	0.27388	−2.33784**	−3.39120**	−3.60866**
CG31864	***FBgn0051864***	FBtr0080288	5	372	0.00369	0.00730	0.505	0.23832	−0.75199	−0.41017	−0.48709
CG12592	***FBgn0037811***	FBtr0082233	18	495	0.00113	0.00208	0.543	0.10899	−2.42979***	−3.51295**	−3.68453**
CG31687	***FBgn0051687***	FBtr0302224	15	1056	0.00248	0.00489	0.507	0.08065	−2.35146***	−3.22751**	−3.43465**
CG30457	***FBgn0050457***	FBtr0086998	31	570	0.00380	0.03389	0.112	0.00211**	−1.45689	−2.22966*	−2.33216*
CG17196	***FBgn0039368***	FBtr0084922	29	831	0.00337	0.01432	0.235	0.00322**	−1.92071*	−1.55541	−1.97526

[a] 基因名称：CG ID 和 FBgn ID 分别是一个基因的两种不同名称；FBtr ID 是指一个基因的转录本名称。[b] d_N 指每个错义替换位点上错义替换的平均数目；d_S 指每个同义替换位点上同义替换的平均数目。[c] Z 检验的 *P* 值。

[a] Gene ID: CG ID and FBgn ID denote two names of each new gene or parental gene; [b] d_N: Mean number of nucleotide substitution per non-synonymous site, d_S: Mean number of nucleotide substitution per synonymous site. [c] *P* value for Z test.

*: *P*<0.05, **: *P*<0.01; ***: *P*<0.001 (Z test and Tajima’s D Test).

*: 0.10>*P* > 0.05, **: *P*<0.02; ***: *P*<(Fu and Li’s D*/F* Test).

表 4　嵌合新基因与其祖先基因的表达组织

Tab. 4　Expressed tissues of chimeric new genes and their parental genes

新基因 New gene	表达组织 Exp tiss [a]	祖先基因 A Par gene A [b]	表达组织 Exp tiss [a]	祖先基因 B Par gene B [b]	表达组织 Exp tiss [a]
CG31904	H	*CG13796*	E	*CG7216*	H
CG18853	UN	*CG12822*	UN	*CG11205*	AT,E
CG32318	UN	*CG9191*	AT,E,O,SL	*CG9187*	EG
CG31864	AT	*CG12264*	AT,E,H,LP,O,SL	*CG5202*	SL
CG12592	UN	*CG18545*	UN	*CG12819*	AT,E,LP,O,SL
CG31687	UN	*CG2508*	AT,E	*CG31688*	E,SL
CR18217	E	*CG17286*	AT,E,LP,O,SL	*CG4098*	E,SL
CG31668	UN	*CG33124*	L	*CG8451*	AT,E,H,LP,O
CG6653	E	*CG31002*	E	*CG17200*	UN
CG17196	UN	*CG17197*	UN	*CG17195*	UN
CG30457	E	*CG10953*	E	*CG13705*	E,H
CG11961	E,EG,H,LP	*CG9416*	E,O	*CG30049*	UN
CG3978	E	*CG9656*	E	*CG10278*	E,LP
CG6844	UN	*CG5610*	UN	*CG11348*	B,E,H

[a] Exp tiss 表示新基因或者祖先基因表达的组织。[b] Par gene A 表示祖先基因 A；Par gene B 表示祖先基因 B；名称简写——E：胚胎；EG：胚胎生殖腺； L：幼虫；LP：幼虫-早期；AT：成虫精巢；H：头； SL：Schneider L2 细胞；O：卵；B：脑； UN：未知。

[a] Exp tiss denotes the tissues in which new genes or parental genes are expressed; [b] Par gene A denotes parental gene A, while Par gene B denotes parental gene B; The abbreviation of the tissues, organs or cell lines (E: embryo; EG: embryonic gonads; L: larvae; LP: L-early pupae; AT: adult testes; H: head; SL: Schneider L2 cell line; O: ovary; B: brain; UN: unknown).

胚胎、幼虫、蛹、雄性成虫与雌性成虫 5 种虫态中，14 个嵌合基因中有 11 个(78.6%)在 3 种及以上的时期呈现中等及以上的表达(图 2)，这提示大部分嵌合基因在多种组织中表达并可能执行相关的功能。可能发生假基因化的基因也表现出多样化的表达。可能发生假基因化的 8 个基因中有 6 个在 3 种及以上的时期呈现中等及以上的表达。其中，*CG18853* 在所有的 30 个时期中都维持中等以上的表达水平。*CG11961* 在除胚胎 00～02 h 以外的时期中都维持中等以上的表达水平，提示它可能是在合子形成后表达的基因。在 14 个嵌合基因中有 6 个(42.9%)在胚胎 00-02 h 内呈现中等以上的表达水平，这提示它们有可能是母源性的表达。在这 6 个基因中 *CG12592* 在胚胎发育的 24 h 内只在 00～02 h 内呈现中等以上的表达水平(图 2)，这提示它可能是个母源效应基因。

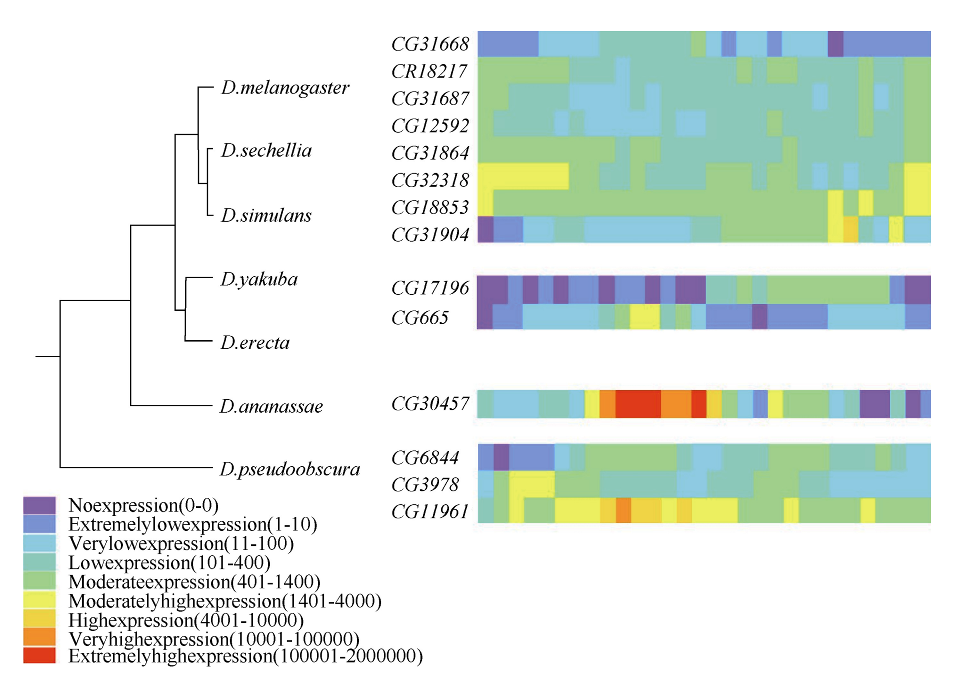


图 2 嵌合基因的表达模式

Fig. 2 Expression patterns of young chimeric genes

Drosophila melanogaster 亚群物种的系统发育树如左，对应的嵌合基因以及其表达模式如右。

从左往右，这些表达数据分别来源于 30 个不同的发育时期，即胚胎 00～02 h、02～04 h、04～06 h、06～08 h、08～10 h、10～12 h、12～14 h、14～16 h、16～18 h、18～20 h、20～22 h、22～24 h、幼虫 L1、L2、L3 12 h、L3 膨胀期(puffstage) 1～2、L3 膨胀期 3～6、L3 膨胀期 7～9、白色预蛹(white prepupae)前期、12 h、24 h、蛹 2 d、3 d、4 d、成虫雄性 1 d、5 d、30 d、成虫雌性 1 d、5 d、30 d。我们将这些发育时期分为 5 种，即胚胎、幼虫、蛹、成虫雄性和成虫雌性。表达量高低分为 9 个级别，即不表达(0)、极度低表达(1～-10)、表达量很低(11～100)、低表达(101～400)、中等表达(401～1400)、中等高度表达(1401～4000)、高表达(4001～1 万)、表达量很高(1.0001 万～10 万)、极度高表达(10.0001 万～200 万)。

The phylogenetic tree of *Drosophila melanogaster* subgroup was shown on the left, while the corresponding chimeric genes and their expression patterns were shown on the right.

From left to right, these expression data were derived from 30 developmental stages, including embryo 00-02 h, 02-04 h, 04-06 h, 06-08 h, 08-10 h, 10-12 h, 12-14 h, 14-16 h, 16-18 h, 18-20 h, 20-22 h, 22-24 h, larva L1, L2, L3 12h old, L3 puffstage 1-2 h, L3 puffstage 3-6 h, L3 puffstage 7-9 h, white prepupae new, white prepupae 12 h, 24 h, pupae 2 d postWPP, pupae 3 d postWPP, pupae 4 d postWPP, adult male 01 d, 05 d, 30 d, adult female 01 d , 05 d, 30 d. The expression levels could be divided to 9 levels, including no expression (0-0), extremely low expression (1-10), very low expression (11-100), low expression (101-400), moderate expression (401-1400), moderately high expression (1401-4000), high expression (4001-10000), very high expression (10001-100000), extremely high expression (100001-2000000).

在成虫的两种性别中，6 个(42.9%)嵌合基因在两种性别中都呈现中等以上的表达水平，4 个(28.6%)基因在两种性别中都不表达或者表达量非常低，2 个(14.3%)基因(*CG32318* 和 *CG31864*)仅在雌性中呈现中等以上的表达水平，2 个(14.3%)基因(*CG17196* 和 *CG6844*)仅在雄性中呈现中等以上的表达水平。其中 *CG17196* 在雌性以及胚胎和幼虫早期表达量极低或者不表达，在幼虫晚期和蛹以及成虫雄性中高表达，因此它们可能跟精巢的发育有关。此前的报道显示，逆转座形成的新基因往往倾向于显现雄性专一性或者精巢专一性表达模式(Betrán et al, 2002; Emerson et al, 2004)。我们的结果表明，与逆转座形成的新基因不一样，DNA 水平外显子重排形成的嵌合基因呈现时空多样性的表达模式，而不限于精巢或者雄性专一性表达。因此，如果它们有功能，则可能具有更为多样性的生物学功能。

3 讨 论

3.1 57.1%的嵌合基因可能正在变成假基因或者非蛋白质编码的 RNA 基因

我们对 *D. melanogaster* 及其亚群中的嵌合新基因的研究表明，14 个嵌合基因中有 8 个(57.1%)积累了提前终止子突变，因此它们可能是假基因，也有可能如 *sphinx* 一样变成了或者正在变成非编码 RNA (non-coding RNA，ncRNA)基因。其中在 *D. melanogaster* 特有的 8 个嵌合基因中有 4 个积累了提前终止子的核苷酸替换(表 2, 3)。在果蝇其他物种

中保留下来的 6 个新基因中有 4 个积累了提前终止子突变。这些结果显示，新起源的嵌合基因中绝大多数会积累有害突变破坏读码框而假基因化，在基因组上被保留下来的嵌合基因依然会被假基因化。因此，能在基因组上长期保留并执行生物学功能的嵌合基因仅占很小的比例。此前关于嵌合新基因的研究显示，嵌合新基因在 *D. melanogaster* 及其亚群中以每百万年 11.4 个新基因的速度产生，其中能在基因组上被保留下来的，仅占 1.4% (Rogers et al, 2009)。虽然在 14 个嵌合基因中只有 *CR18217* 被注释成假基因，但群体遗传学分析显示，57.1%的嵌合基因在群体中积累了提前终止子突变或者在在选择上呈现中性，这提示大部分嵌合新基因可能正在假基因化或者正在变成非蛋白质编码的 RNA 基因。

Kimura (1983)提出了一个经典的模型—等待模型(waiting model)——来描述重复基因(产生新基因最重要的方式之一)如何获得新功能并在基因组上最终保留下来。基因重复后，作用于一个或者两个拷贝上的负选择得到放松，因而能积累中性突变，甚至有害突变；最终，一个或者两个拷贝上积累的部分突变要么被负选择所清除，要么被正选择所保留(Kimura, 1983)。按照这个模型，新基因起源后的早期阶段，由于负选择放松，新基因往往表现为选择上的中性状态或者近似中性状态，以更快地积累各种突变，从而为新功能的进化积累序列上的材料。在这个早期阶段，新基因往往表现为假基因的特征。因此，假基因化可能是新基因进化出新功能的一个中间状态。而正在假基因化的嵌合基因有多少能进化出新功能也有待进一步分析。

CR18217 曾经被 FlyBase 网站(http://flybase.org/)注释成蛋白质编码基因 *CG18217* (基因组版本：Dmel r 4.3, FB2006_10)，现在被注释成假基因，且基因结构与以前不一样(图 3A)。我们反转录 PCR(RT-PCR)与测序的结果也支持新的基因结构，即早期版本 *CG18217* 中注释的第 3 个内含子是转录的。此外，我们调查了 *CR18217* 在 *D. melanogaster* 不同群体中的序列，发现其以前注释的 ORF 内存在 3 个移码突变(frame shift mutations) (3 个外显子上分别有 1、31、4 bp 的缺失),它的蛋白质编码能力在群体中并没有被固定下来。因此，*CR18217* 很可能的确是假基因。最近释放的转录组测序的数据也支持现在的基因结构，转录组测序产生

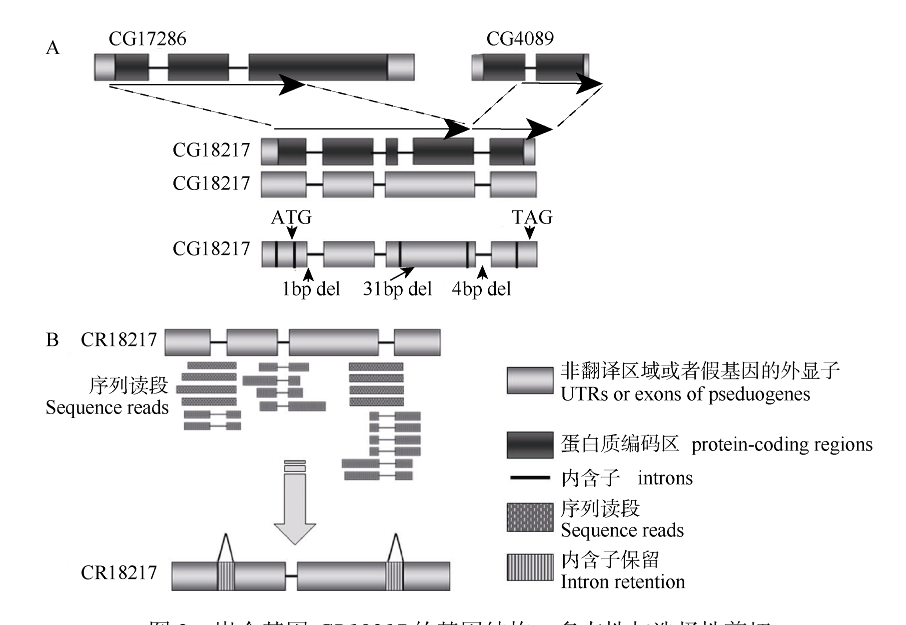


图 3　嵌合基因 *CR18217* 的基因结构、多态性与选择性剪切

Fig. 3　Gene structure, polymorphsims and alternative splicing of chimeric genes *CR18217*

A：*CR18217* 由两个祖先基因 *CG17286* 和 *CG4098* 通过外显子重排而形成，其最初被注释成蛋白质编码基因 *CG18217*, RT-PCR 证实现在注释的基因结构是正确的, *Drosophila melanogaster* 群体分析显示它有三个缺失(del)多态性破坏了此前注释的读码框，因而 *CR18217* 的确不是蛋白质编码基因。B：转录组测序产生的序列读段显示其存在多种选择性剪切模式，其中第一个和最后一个内含子都是选择性剪切的内含子。

A: *CR18217* arose through fusion of parts of two parental genes *CG17286* and *CG4098* by exon shuffling. *CR18217* was initially annotated as a protein-coding gene. RT-PCR results showed the current gene structure is correct. *Drosophila melanogaster* population analysis showed that two deletion (del) polymorphisms disrupted its initially annotated ORF (open reading frame), thus *CR18217* is indeed not a protein-coding gene. B: the sequence reads generated in transcriptome sequencing showed that *CR18217* has multiple alternatively spliced isoforms.

的序列读段能比对到*CR18217*上，且所有读段都显示第二个内含子被剪切掉，但第一个和最后一个内含子在不同的读段中被选择性剪切或者保留(图3B)。因此，*CR18217* 可能存在多种不同的选择剪切模式。

除此之外，在14个嵌合基因中还有6个存在多种选择性剪切形式。其中仅有2个包涵在新近起源的8个嵌合基因中，这说明新近起源的嵌合基因中只有少数(25%, 2/8)存在选择性剪切；而起源较早的嵌合基因中大部分(66.7%, 4/6)存在选择性剪切。因此，选择性剪切或可以成为检测嵌合基因能否在基因组上保留下来的一个标志，因为新基因进化出选择性剪切很可能是为了功能进化的需要，即它们正在往新功能进化的道路上前进。*CG31687*受到显著的负选择(表3),很可能有功能。尽管*CG31904*和*CR18217*的蛋白质编码能力在群体中并未被固定下来，但其存在多种选择性剪切模式，与此前报道的*sphinx*相似(Wang et al, 2002)，它们可能是有功能的非蛋白质编码 RNA 基因；或者处于新基因进化的中间状态，与等待模型一致，在未来它们会进化出新的 ORF 并清除有害突变而维持新的蛋白质编码能力和相应的功能。在基因组上保留了较长时间的*CG11961*、*CG3978*和*CG6844*在*D. melanogaster*不同群体中也有提前终止子突变，它们与*CG31904*和 *CR18217* 一样也可能存在两种进化命运——功能的非蛋白质编码 RNA 基因和进化出新的蛋白质编码能力。

综上所述，DNA水平外显子重排而产生的嵌合基因，在群体中容易积累有害突变而假基因化，如发生提前终止子突变。它们在进化中以一个非常低的比例保留下来。尽管这些嵌合基因大部分都将会以假基因化的方式退出进化舞台，但仍有少数因为具有新的表达模式和/或选择性剪切模式。因此，它们可能如同*sphinx* (Wang et al, 2002)一样，是具有功能的非蛋白质编码的 RNA 基因；或者如等待模型(Kimura, 1983)所描述的一样，由于选择放松而积累中性，甚至有害突变。在未来的进化中，它们将进化出新的开放读码框，并进化出新的蛋白质功能。因此，假基因化很可能是这部分嵌合基因进化中的中间状态。

3.2 外显子重排产生的嵌合基因具有更为多样化的表达

嵌合基因主要由RNA介导外显子重排(逆转座)和 DNA 水平的外显子重排两种机制而产生。逆转座形成的新基因，包括嵌合新基因，往往表现为雄性特异性或者精巢特异性表达，且整体上倾向于逃离X染色体，这可能是为了逃避雄性减数分裂中X染色体失活(MSCI: meiotic sex chromosome inactivation) (Kaessmann et al, 2009)。

与逆转座形成的嵌合基因相比，关于由两个或者多个基因通过 DNA 水平外显子重排而形成的嵌合基因的报道相对较少，其系统性研究也不多。我们在黑腹果蝇及其亚群中研究了由两个基因通过外显子重排而产生的嵌合基因的表达，发现这些嵌合基因，包括可能正在发生假基因化的基因，并不倾向于在雄性或者精巢中专一性表达，而是在两种性别和多个发育时期中都表达，其表达范围更为广泛(图2，表4)。这说明由外显子重排而形成的嵌合基因的生物学功能并不局限于雄性或者精巢中，更可能具有更为广泛的生物学功能。

致谢：感谢本实验室的任娟同学在实验中给予的帮助，感谢比较基因组学小组在测序上提供的支持和帮助，感谢本实验室的李学燕、相辉、李昕和丁昀博士在论文修改给予的富有建设性的意见。

参考文献：

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs [J]. *Nucleic Acids Res,* **25**(17): 3389-3402.

Babushok DV, Ohshima K, Ostertag EM, Chen X, Wang Y, Mandal PK, Okada N, Abrams CS, Kazazian HH Jr. 2007. A novel testis ubiquitin-binding protein gene arose by exon shuffling in hominoids [J]. *Genome Res,* **17**(8): 1129-1138.

Bai Y, Casola C, Feschotte C, Betrán E. 2007. Comparative genomics reveals a constant rate of origination and convergent acquisition of functional retrogenes in *Drosophila* [J]. *Genome Biol,* **8**(1): R11.

Betrán E, Thornton K, Long M. 2002. Retroposed new genes out of the X in Drosophila [J]. *Genome Res,* **12**(12): 1854-1859.

Brennan G, Kozyrev Y, Hu SL. 2008. TRIMCyp expression in Old World primates Macaca nemestrina and Macaca fascicularis [J]. *Proc Natl Acad Sci U S A,* **105**(9): 3569-3574.

Courseaux A, Nahon JL. 2001. Birth of two chimeric genes in the Hominidae lineage [J]. *Science,* **291**(5507): 1293-1297.

Emerson JJ, Kaessmann H, Betrán E, Long M. 2004. Extensive gene traffic on the mammalian X chromosome [J]. *Science,* **303**(5657): 537-540.

Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J. 1999. Preservation of duplicate genes by complementary, degenerative mutations [J]. *Genetics,* **151**(4): 1531-1545.

Fu YX. 1997. Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection [J]. *Genetics,* **147**: 915-925.

Fu YX, Li WH. 1993. Statistical tests of neutrality of mutations [J]. *Genetics,* **133**(3): 693-709.

Gilbert W. 1987. The exon theory of genes [J]. *Cold Spring Harbor Symp Quant Biol*, **52**: 901-905.

Jones CD, Begun DJ. 2005. Parallel evolution of chimeric fusion genes [J]. *Proc Natl Acad Sci U S A,* **102**(32): 11373-11378.

Kaessmann H, Vinckenbosch N, Long M. 2009. RNA-based gene duplication: mechanistic and evolutionary insights [J]. *Nat Rev Genet,* **10**(1): 19-31.

Kimura M. 1983. The Neutral Theory of Molecular Evolution [M]. Cambridge: Cambridge Univ Press.

Li X, Yang S, Peng LX, Wang W. 2004. Origin and evolution of new genes [J]. *Chn Sci Bull,* **49**(13): 1219-1225. [李昕, 杨爽, 彭立新, 王文. 2004. 新基因的起源和进化. 科学通报, **49**(13): 1219-1225.]

Long M, Betrán E, Thornton K, Wang W. 2003. The origin of new genes: glimpses from the young and old [J]. *Nat Rev Genet,* **4**(11): 865-875.

Long M, Langley CH. 1993. Natural selection and the origin of jingwei, a chimeric processed functional gene in *Drosophila* [J]. *Science,* **260**(5104): 91-95.

Long M, Wang W, Zhang J. 1999. Origin of new genes and source for N-terminal domain of the chimerical gene, jingwei, in *Drosophila* [J]. Gene, **238**(1): 135-141.

Lynch M, Conery JS. 2000. The evolutionary fate and consequences of duplicate genes [J]. *Science,* **290**(5494): 1151-1155.

Marques AC, Dupanloup I, Vinckenbosch N, Reymond A, Kaessmann H. 2005. Emergence of young human genes after a burst of retroposition in primates [J]. *PLoS Biol,* **3**(11): e357.

Rogers RL, Bedford T, Hartl DL. 2009. Formation and longevity of chimeric and duplicate genes in *Drosophila melanogaster* [J]. *Genetics,* **181**(1): 313-322.

Rohozinski J, Lamb DJ, Bishop CE. 2006. UTP14c is a recently acquired retrogene associated with spermatogenesis and fertility in man [J]. *Biol Reprod,* **74**(4): 644-651.

Rozas J, Sánchez-DelBarrio JC, Messeguer X, Rozas R. 2003. DnaSP, DNA polymorphism analyses by the coalescent and other methods [J]. *Bioinformatics,* **19**(18): 2496-2497.

Sayah DM, Sokolskaja E, Berthoux L, Luban J. 2004. Cyclophilin A retrotransposition into TRIM5 explains owl monkey resistance to HIV-1 [J]. *Nature,* **430**(6999): 569-573.

Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism [J]. *Genetics,* **123**(3): 585-595.

Tamura K, Dudley J, Nei M, Kumar S. 2007. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0 [J]. *Mol Biol Evol,* **24**(8): 1596-1599.

van Rijk AA, de Jong WW, Bloemendal H. 1999. Exon shuffling mimicked in cell culture [J]. *Proc Natl Acad Sci U S A,* **96**(14): 8074-8079.

Virgen CA, Kratovac Z, Bieniasz PD, Hatziioannou T. 2008. Independent genesis of chimeric TRIM5-cyclophilin proteins in two primate species [J]. *Proc Natl Acad Sci U S A,* **105**(9): 3563-3568.

Wang W, Brunet FG, Nevo E, Long M. 2002. Origin of sphinx, a young chimeric RNA gene in *Drosophila melanogaster* [J]. *Proc Natl Acad Sci U S A,* **99**(7): 4448-4453.

Wang W, Zhang J, Alvarez C, Llopart A, Long M. 2000. The origin of the Jingwei gene and the complex modular structure of its parental gene, yellow emperor, in *Drosophila melanogaster* [J]. *Mol Biol Evol,* **17**(9): 1294-1301.

Wang W, Zheng H, Fan C, Li J, Shi J, Cai Z, Zhang G, Liu D, Zhang J, Vang S, Lu Z, Wong GK, Long M, Wang J. 2006. High rate of chimeric gene origination by retroposition in plant genomes [J]. *Plant Cell,* **18**(8): 1791-1802.

Wilson SJ, Webb BL, Ylinen LM, Verschoor E, Heeney JL, Towers GJ. 2008. Independent evolution of an antiviral TRIMCyp in rhesus macaques [J]. *Proc Natl Acad Sci U S A,* **105**(9): 3557-3562.

Zhou Q, Wang W. 2008. On the origin and evolution of new genes－a genomic and experimental perspective [J]. *J Genet Genomics,* **35**(11): 639-648.

Zhou Q, Zhang G, Zhang Y, Xu S, Zhao R, Zhan Z, Li X, Ding Y, Yang S, Wang W. 2008. On the origin of new genes in *Drosophila* [J]. *Genome Res,* **18**(9): 1446-1455.